IJDC | *Peer-Reviewed Paper*

What Factors Influence Where Researchers Deposit their Data? A Survey of Researcher Submissions to Data Repositories

Shea Swauger
University Libraries
Colorado State University

Todd J. Vision
Department of Biology
University of North Carolina at Chapel Hill

Abstract

In order to better understand the factors that most influence where researchers deposit their data when they have a choice, we collected survey data from researchers who deposited phylogenetic data in either the TreeBASE or Dryad data repositories. Respondents were asked to rank the relative importance of eight possible factors. We found that factors differed in importance for both TreeBASE and Dryad, and that the rankings differed subtly but significantly between TreeBASE and Dryad users. On average, TreeBASE users ranked the domain specialization of the repository highest, while Dryad users ranked as equal highest their trust in the persistence of the repository and the ease of its data submission process. Interestingly, respondents (particularly Dryad users) were strongly divided as to whether being directed to choose a particular repository by a journal policy or funding agency was among the most or least important factors. Some users reported depositing their data in multiple repositories and archiving their data voluntarily.

Received 4 November 2013 | *Revision received* 5 January 2015 | *Accepted* 5 January 2015

Correspondence should be addressed to Shea Swauger, 1019 Campus Delivery Fort Collins, CO 80523. Email: shea.swauger@colostate.edu

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

The factors that affect where researchers deposit their data are not widely understood. Yet, research institutions, journals, grant funders and repositories have policies and workflows that are intended to directly affect research data deposition. As the number of options for data repositories grow, researchers and research support organizations must weigh the different features that repositories offer and decide which ones to use or to promote. Knowing what features are of importance to researchers will help repositories to design more useful services and help research support organizations create better informed policies. This motivated us to survey researchers working with a particular datatype, who have submitted data to one of two alternative repositories, in order to better understand what factors drove their choice.

Literature Review

We first reviewed the literature on factors that may be relevant for understanding user choice. Previous researchers have identified a number of factors that commonly differ between data repositories that may influence where researchers choose to deposit their data. Here, we discuss eight factors in particular that could be relevant in the repository comparison explored here. These are summarized in Table 1.

Specialization

The specialization of the repository for a particular data type may influence where researchers deposit their data. A homogenous collection of data (e.g. data type, structure or format) facilitates the use of tools that can search and manipulate data in ways that are more difficult to achieve with heterogeneous collections. For example, some of the popularity of the GenBank repository, a specialized repository for genetic sequence data, can likely be ascribed to the availability of powerful discovery and analysis tools that can process all the content in the collection (e.g. Altschul et al., 1990). Researchers may be more inclined to deposit their data into a repository that can accommodate such tools.

Prestige

In their evaluation of the perception of journal prestige, Catling et al. (2009) found that a journal's impact factor, visibility within a discipline and selectivity all contribute to its level of prestige. In a similar way, researchers may perceive differences in prestige among repositories. Such perceptions, influenced for example by the use of a repository by a researcher's peers, might affect where a researcher chooses to deposit their data.

Ease

The difficulty or ease of the data submission process between repositories can vary depending on submission workflow, format requirements, website usability, metadata requirements and the amount of supplementary information required per data deposit. In their survey of the usability of software repositories, Clayton et al. (2000) were repeatedly

surprised at how difficult it was to navigate the repositories they evaluated and how much longer it took to complete tasks than they expected. In a survey of university professors and faculty, Jacobs and Winslow (2004) found that respondents felt overburdened and did not have sufficient time to complete all of the tasks for which they were responsible. Impatience may be exacerbated when data archiving is optional, since researchers resent being burdened with professional obligations, including repository depositions, which are seen to be outside of their normal duties (Fried Foster and Gibbon, 2005). Considering these factors, the ease of the data submission process, and thus the amount of time it takes to submit data to a repository, appears to be an important feature for researchers when choosing a repository.

Metadata

Countering the above factor, there might be a trade off between the ease of submission and the quality of a repository's metadata. In their paper describing the value of metadata for ecological data, Fergraus et al. (2005) argue for more rigorous and descriptive metadata practices in the ecological data community in order to increase the usability and long-term value of collected data. Goovaerts and Leinders (2012) contend that rich metadata allows for greater accessibility and superior services that can be offered to repository end-users. Berkely et al. (2009) claim that as the amount of data available to researchers continues to grow, metadata will become all the more important to be able to locate and interpret that data. Some repositories require richer metadata at the time of submission or employ curation staff that enrich the user-provided metadata. If researchers place importance on their data being reusable, they may choose to archive their data in a repository that has features promoting higher quality metadata.

Trust

Researchers may perceive one repository to be more stable than another based on how long they have existed, their funding models or their participation in a digital preservation system. In their overview of preservation initiatives, Bone and Burns (2011) present several content perpetuity systems that libraries, archives and repositories can use to guarantee the digital information they store will be accessible if they discontinue their services. ISO 16363, sometimes called CCSDS 652.0-M-1, is a standard that is used to measure the trustworthiness of digital repositories and outlines the attributes repositories must have in order to meet its certification requirements (ISO, 2012). Such standards have raised awareness in the data community as to the importance of a repository's trustworthiness. Standard criteria are emerging for journal/publishers to use in deciding what repositories are acceptable or preferred (Callaghan et al., 2014). If the persistence of a repository and its ability to safeguard its digital content are important factors to researchers, they could influence where they choose to deposit their data.

Credit

Repositories may differ in the extent to which they support researchers seeking scholarly credit for their contributions, for instance by supporting data citation and usage tracking. Many have suggested that proper data citation, made possible in part by the use of

persistent identifiers, will help to incentivize data archiving by allowing data producers to receive credit for their data (Edmunds et al., 2012; CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013; Costello et al., 2013). This may mean that repositories that assign persistent identifiers to data, thus allowing easy citation by others, may be more attractive to researchers when deciding where to archive their data.

Limit Reuse

Repositories may differ in the extent to which they allow data submitters to control the reuse of the data by, for example, restricting public access for a period of time. In their survey of geneticists, Campbell et al. (2002) found that of those respondents who had denied fellow academics access to the data of their published article, 53% did so in order to protect their own ability to publish subsequent articles. Thus, researchers may choose a repository that allows them to restrict the terms of reuse or limit public access for a period of time.

Directed by Journal

Being directed to choose a repository by a research funder, journal or institution may be an important factor in deciding where to archive one's data. While many journals only recommend or require that researchers archive their data in a public repository, an increasing number recommend or require particular data repositories (e.g. Magee et al., 2014). For example, Whitlock (2011), in discussing best practices with regards to recently adopted journal data archiving policies in ecology and evolution, recommends specific repositories for different data types:

> "Choose an archive that is most suitable for your type of data. For example, GenBank is of course the right place for DNA sequence data; TreeBASE is the right place for phylogenetic trees and the data matrices used to generate them; and archives such as GEO support microarray, next generation sequencing and other forms of high-throughput functional genomic data. Other data have multiple possible hosts. All data in the fields of ecology and evolutionary biology can be archived at the Dryad repository or KNB, provided there is not an established site for that kind of data" (Whitlock, 2011).

Funding agencies are also increasingly adopting data sharing policies that encourage the deposition of research data into particular repositories (Jones, 2012) which, depending on the policy, may be run by the funding agency, the researcher's institution, or another organization. Institutional policies, while still uncommon, may also be a relevant factor at some institutions (e.g. Rice et al., 2013).

Methods

We used an anonymous survey to measure that factors were perceived by researchers as more or less important in choosing where to archive data. Our sample was drawn from researchers who had archived phylogenetic trees in one of two repositories frequently

 doi:10.2218/ijdc.v10i1.289

Table 1. Labels for the factors considered in this study.

Factor	Label
Specialization of the repository for your data	Specialization
Prestige of the repository	Prestige
Ease of the data submission process	Ease
Extent of metadata quality control (by the researcher or repository curators)	Metadata
Trust in the persistence of the repository	Trust
Policies of the repository that promote scholarly credit (e.g. assigning DOIs for data citation)	Credit
Policies of the repository that limit reuse by others (licenses, embargoes)	Limit Reuse
Directed to choose the repository by your research funder, journal or institution	Directed by Journal

used for phylogenetic data: TreeBASE[1] and Dryad[2]. The survey asked respondents to rank the importance of the eight factors listed in Table 1 in their choice of repository (with one being the most important, and eight being the least).

A comparison of TreeBASE and Dryad is informative because, while these are the main repositories for phylogenetic data associated with published studies (Stoltzfus et al., 2012), they offer different features and services that may be important to different user groups and therefore affect where users choose to deposit their data. For example, TreeBASE only accepts phylogenetic data, requires that data to be deposited in specific formats and relies solely on the metadata provided by their users. In contrast, Dryad accepts any non-sensitive scientific or medical data in practically any format, providing that the depositor follows general guidelines of reasonable data description and provides formats that are accessible to end users. Dryad also employs curation staff that perform quality control and enhance user metadata. As a result of these differences, we hypothesized that TreeBASE and Dryad users would rank the importance of the factors differently, for instance with specialization being more important to TreeBASE users and ease being more important to Dryad users.

The survey population consisted of all users that had submitted phylogenetic trees to either TreeBASE or Dryad from 2010 through 2013. For Dryad, we searched for data packages using a set of case-insensitive keywords (including 'phylogeny', 'phylogenetic', 'tree', 'nexus', and 'taxa'). From these results, we inspected the data packages to verify that they included a phylogenetic tree and obtained the email addresses of the depositors. We emailed an invitation to complete the survey to each user, giving them 14 days to respond with a reminder after seven days.

The content of our emails, survey and the nature of our study were approved by the

[1] TreeBASE: http://treebase.org/
[2] Dryad: http://datadryad.org/

Institutional Review Board at the University of North Carolina at Chapel Hill (IRB study 13-1039).

We compared the responses to those expected under simple null hypotheses using two different statistics, following Brokhoff et al. (2003). Friedman's statistic, F, was used to test whether the mean rank of each factor was significantly different from that expected by chance. It was calculated as follows, where $\text{obs}(i, j)$ is the number of respondents that assigned rank $j = \{1..t\}$ to factor $i = \{1..t\}$ and n is the total number of respondents.

$$R_i = \sum_{j=1}^{t} j \cdot \text{obs}(i, j)$$

$$F = \frac{12}{nt(t+1)} \sum_{i}^{t} \left[R_i - \frac{n(t+1)}{2} \right]^2$$

This was tested for significance against a χ^2 distribution with $t - 1 = 7$ degrees of freedom (d.f.).

Anderson's statistic, A, was used to test whether the overall distribution of ranks was significantly different from that expected by chance. It was calculated as follows, where $\exp(i, j)$ is the expected number of respondents who assigned rank j to factor i.

$$A = \frac{t-1}{t} \sum_{i,j} \frac{[\text{obs}(i, j) - \exp(i, j)]^2}{\exp(i, j)}$$

This is tested against a χ^2 distribution with $(t - 1)^2 = 49$ d.f.

The expected values were derived in two different ways. In testing for the equality of distributions within a repository, $\exp(i, j) = n/t$. For testing whether the preferences of Dryad users were the same as those for TreeBASE users, the expected values were instead calculated as follows, where subscripts D and T denote responses from Dryad and TreeBASE users, respectively.

$$\exp_D(i, j) = \frac{n_D}{n_T} \text{obs}_T(i, j)$$

Rank-factor combinations for which $\exp_D(i, j) = 0$ were not included in the calculation of A.

Additional questions were asked to aid in qualitative interpretation. Respondents could list other factors in a free-text response. They were asked if they deposit their data in more than one repository and if so, which ones and under what circumstances. Lastly, the survey asked if respondents had a repository at their institution and if so, why they do or do not choose to use it.

Results

In total, we sent 819 surveys and received 146 responses (a 17.8% response rate); 651 surveys went to TreeBASE users with 109 responding (16.7%); 125 surveys went to Dryad users with 31 responding (24.8%). Of all the respondents who began the survey, six did not complete it and 43 of the email invitations sent to respondents were returned as a failed delivery, typically indicating that their email addresses were no longer in

 doi:10.2218/ijdc.v10i1.289

Table 2. Percentage of respondents that assigned a given rank to each factor among users of (*A*) TreeBASE ($n_T = 109$) and (*B*) Dryad ($n_D = 31$).

	A. TreeBASE							
Factor	1st	2nd	3rd	4th	5th	6th	7th	8th
Specialization	22.0	24.8	21.1	13.8	8.3	5.5	1.8	2.8
Prestige	6.4	12.8	12.8	13.8	14.7	8.3	9.2	22.0
Ease	12.8	21.1	27.5	16.5	10.1	5.5	4.6	1.8
Metadata	0.9	1.8	9.2	18.3	29.4	26.6	9.2	4.6
Trust	16.5	20.2	20.2	14.7	13.8	8.3	5.5	0.9
Credit	0.0	5.5	2.8	12.8	11.9	29.4	27.5	10.1
Limit Reuse	0.0	0.9	0.0	5.5	6.4	10.1	32.1	45.0
Directed by Journal	41.3	12.8	6.4	4.6	5.5	6.4	10.1	12.8

	B. Dryad							
Factor	1st	2nd	3rd	4th	5th	6th	7th	8th
Specialization	22.6	6.5	22.6	16.1	16.1	3.2	12.9	0.0
Prestige	3.2	16.1	6.5	12.9	12.9	19.4	16.1	12.9
Ease	19.4	19.4	25.8	22.6	3.2	6.5	3.2	0.0
Metadata	0.0	0.0	9.7	16.1	35.5	19.4	3.2	16.1
Trust	22.6	22.6	9.7	16.1	9.7	16.1	3.2	0.0
Credit	3.2	16.1	19.4	12.9	9.7	6.5	29.0	3.2
Limit Reuse	0.0	3.2	0.0	0.0	3.2	22.6	29.0	41.9
Directed by Journal	29.0	16.1	6.5	3.2	9.7	6.5	3.2	25.8

use. Table 2 summarizes the frequency with which each factor was assigned each rank separately for TreeBASE and Dryad users.

We were able to reject the null hypothesis that the mean ranks of the factors did not differ among the factors for both TreeBASE (Friedman's test: $F = 338.15$, 7 d.f., $p < 0.001$) and Dryad ($F = 78.62$, 7 d.f., $p < 0.001$) users. We could also reject the null hypothesis that the distributions of the ranks did not differ among the factors for both TreeBASE (Anderson's test: $A = 132.10$, 49 d.f., $p < 0.001$) and Dryad ($A = 106.06$, 49 d.f., $p < 0.001$). Note, however, that the small number of Dryad users necessitates that the results of Anderson's tests need to be interpreted with caution for that group.

We then used Anderson's statistic to test if the distributions of ranks were identical between TreeBASE and Dryad users by taking the observed frequencies of ranks from TreeBASE users as the expected frequency for Dryad users. This was rejected ($A = 106.06$, 49 d.f., $p < 0.001$), although the small sample size caveat mentioned above applies here, as well.

Side-by-side comparison of the average and relative rankings between TreeBASE and Dryad users (Table 3) reveals that the four highest ranking factors were the same between

Table 3. Comparison of ranks by users of both repositories.

	TreeBASE		Dryad	
Factor	Mean	Relative Rank	Mean	Relative Rank
Specialization	3.0	1	3.6	3
Ease	3.3	2	3.0	1
Trust	3.4	3.5	3.3	2
Directed by Journal	3.4	3.5	4.1	4
Prestige	4.9	5	5.0	6
Metadata	5.1	6	5.4	7
Credit	5.8	7	4.6	5
Limit Reuse	7.0	8	7.0	8

the two groups of users, but that relative order differs within the highest ranking factors and within the lowest ranking factors. We consider the responses for each factor in turn.

Ease and Specialization

The most important factor for TreeBASE users was Specialization, while for Dryad users it was Ease, although both user groups rated both factors in the top two (TreeBASE) or three (Dryad). No users in either group gave these the lowest rank. Comments from respondents emphasized the importance of both of these factors. TreeBASE users wrote:

- "Special types of data needs special type of repository"
- "I collect so much data and I am so busy as a faculty member that it is important for me to be able to archive my data easily and quickly."
- "Although ease of data submission process is not the most important factor it can be a 'killer' for desire to [deposit] data, resulting in only mandatory submissions being performed."
- "[T]he objective is typically to publish a paper, and to reach that objective as quickly as possible, I followed the publisher's re[q]uirement, and pick[ed] the simplest . . . repository among the ones proposed."

While a Dryad user wrote:

- "Usually the process to upload molecular and associated data is a real pain (for example treebase). Therefore, I believe the key for a successful and widely used repository is to be user friendly and as little time consuming as possible."

Trust

This factor was rated highly by both groups, and in fact higher than Specialization by Dryad users. One TreeBASE user wrote: "Depositing data to ephemeral, or grant-cycle-based databases doesn't ensure long-term data-storage. If you want your manuscript's

data to be relevant decades into the future, database persistence becomes the number one factor." Another TreeBASE user wrote: "If it is not going to persist long-term, why bother?"

Directed by Journal

While this factor had the greatest frequency of being ranked first in both repository use groups, the distribution was in both cases bimodal, with 23% of TreeBASE users and 29% of Dryad users assigning a rank of 7 or 8. A plausible (though untested) explanation for this pattern is that researchers gave high importance to journal instructions when they existed, but that many were publishing in journals that lacked a data policy, or at least lacked one that was explicit about choice of repository. One TreeBASE user wrote: "Quite simply, if a journal wants data to be uploaded to a specific databa[s]e, that is what [I]'ll do in order to publish in that journal." Of the respondents across both groups assigning a rank of 7 or 8, the factors that did rank as most important were Trust (38%), Specialization (26%) and Ease (23%).

While the question allowed for respondents to consider the influence of funder and institutional policies, it is noteworthy that the words 'journal' or 'publish' and their derivatives were mentioned 69 times in the free text answers of both repository user groups, while 'fund' and its derivatives was only mentioned only four times, and 'institution' and its derivatives were mentioned only four times in free-text responses. Furthermore, the latter was only used in the context of institutional repositories rather than institutional policies. Thus, in making repository choices, this sample of researchers seems to be much more aware or concerned with the policies of the journals in which they publish than with the policies of their funders and institutions.

Prestige

Prestige was assigned a wide range of rankings in both user groups. 45% and 50% of TreeBASE and Dryad users, respectively, ranked it among the top four factors. One TreeBASE user wrote: "I've been submitting data to TreeBASE and GenBank for over 20 years. Their longevity and prestige were important considerations" while another, wrote: "I chose the repository that I was most familiar with – not necessarily because of its prestige (I didn't realize repositories had prestige value.)" It is possible that the perceived importance of this factor varies with the researcher's career stage, or with their knowledge of data management practices and repositories.

Metadata

This factor was most commonly assigned moderate to low ranking, with over two thirds of respondents assigning it to rank 4, 5 or 6 in both groups, and it showed relatively little difference between groups.

Credit

The Credit factor had a bimodal distribution for Dryad users with 39% of them ranking it as 1, 2 or 3 (more important) and 39% ranking it as 6, 7 or 8 (less important). However,

TreeBASE users were more uniform in their answers, with 67% ranking it in their bottom three. It would be of interest to understand the reasons for the bimodality among Dryad users, which may be related to the respondent's individual understanding of or attitude toward data citations.

Some user comments suggested that users may not fully separate Prestige, Specialization and Credit, and the unlisted factor of the desire for one's data to be seen, cited and/or reused. One TreeBASE respondent wrote: "I put my phylogeny in TreeBase because it is widely known and thus I hope that my phylogeny will be found by and be useful for the greatest number of other researchers." Another wrote: "We used GenBank, which is the standard repository for plant systematics, my field of research ... GenBank is the first place that anyone in the field looks for sequence data," and a third, wrote: "[A repository's] prestige influences how many people use it."

Limited Reuse

Neither Dryad nor TreeBASE users indicated that policies that limit data reuse were important in deciding where to archive their data. Over 90% of Dryad users and 87% of TreeBASE users ranked it among their bottom three factors, and no respondents ranked it most highly.

In addition to ranking the factors above, three questions on the survey were aimed at measuring the frequency and motivations for depositing data in multiple repositories, and the effect of the availability of an institutional repository (IR) on data archiving habits.

Of the 96 of TreeBASE respondents (88% of the TreeBASE population) who answered the question "Do you deposit your data in multiple repositories? If so, which ones? Under what circumstances do you do this?" 47% indicated they deposited their data into multiple repositories. Of the 23 Dryad respondents who answered the same question (74% of the Dryad population), 56% indicated that they deposited their data into multiple repositories. One consideration for users was the type of data being deposited. For example, one TreeBASE respondent wrote: "I used TreeBase for the phylogenetic matrix as required and Dryad for all the additional supplementary data for the study." Another consideration was the ease of submission; one TreeBASE respondent wrote that they deposit data in multiple repositories "provided that submission is easy!"

When asked if their institution had its own repository that accepts research data, 16% of all respondents responded 'Yes', 62% 'No' and 21% "Don't Know". Those who responded "Yes" were prompted to answer why they did or did not use their IR. Of the 30 responses received, only four indicated that they used their IR, with two of the four stating that their deposits were specimens or samples collected during their research. Only one respondent endorsed his IR, writing: "Why not use it. It is there, easy to use and is an extra safeguard that data is stored for future use." Reasons given (in no particular order) for not using the IR included: (1) it was inappropriate for their kind of data, (2) their IR did not accept data at all, (3) submitters were unfamiliar with how to use it, and (4) that depositing data into an IR was not required for publication and lack of visibility. As one TreeBASE respondent wrote: "I don't use my university's archive because it is not easily accessible, not widely known outside my institution, and not easily searchable."

 doi:10.2218/ijdc.v10i1.289

Discussion and Conclusions

Overall, respondents submitting phylogenetic data to these two repositories rank the factors affecting their choice similarly. The set of factors most important to both repository users in this survey were Specialization, Ease, Trust, and Directed by Journal. Journals appear to be in a particularly influential position for affecting repository choice. Policies directing users to one repository versus another can trump the other factors that would otherwise contribute to the choice of individual researchers. Factors ranked of lower importance to both groups were Prestige, Metadata, Credit and Limit Reuse. The relatively low importance of policies limiting reuse comes as a surprise, because embargos had been seen as critical to the adoption of journal policies mandating archiving in the ecology and evolutionary biology community (Whitlock, 2011) and remain a matter of lively policy debate (Roche et al., 2014). Further work will be needed to determine if this signals a growing level of comfort among researchers with the idea of making data available at the time of publication.

We found significant differences in the distribution of rankings between the two user groups, with the caveat that sample sizes of Dryad users would need to be greater to have greater confidence in the outcome of the test. Some of the differences observed in the relative ranking of factors were consistent with expectations based on differences between the repositories. For instance, users of TreeBASE generally value disciplinary specialization more than ease of submission and the opposite is true for users of Dryad. However, one striking difference for which the interpretation is less clear is the sizeable minority of Dryad users that ranked Credit of moderate importance, in contrast to the TreeBASE users, who uniformly ranked it as being of lesser importance.

While the survey questions mostly focused on disciplinary repositories, it may be possible to apply some lessons to institutional repositories. For one, despite the relatively low ranking of Prestige and Credit, the free-text reasons given for choice of repository do suggest that users want their data to be visible and reused. Furthermore, some researchers expressed a willingness to deposit their data in multiple repositories, provided the submission processes is sufficiently easy. The researchers surveyed here used IRs only rarely. IRs might be able to attract more submissions by focusing on the ease of the deposition process and increasing the visibility of the data they collect. Institutional data policies and support may also still have an influence on the likelihood that research data is publicly archived somewhere, even if it does not affect the choice of repository (Sayogo and Pardo, 2013).

There are a number of limitations to this study and areas for future work. For one, the modest response rates leave open the possibility that the respondents may represent biased samples of Dryad and TreeBASE users. Some factors were not included in this survey that may be relevant, such as user registration policies or deposition fees. Wicherts, Bakker and Molenaar (2011) found that the strength of statistical evidence for the findings in a paper is correlated with the willingness of the researcher to share their data; thus, differences among repositories in the extent of review may also affect repository choice. It is important to recognize that respondents weren't asked to specifically compare TreeBASE and Dryad, nor to report what other repositories they were aware of, which may lead to some mismatch between the stated preferences of the users and the observable differences between these two repositories. A direct comparison

of how different repositories are perceived, how different components of trust are valued (Callaghan et al., 2014), and a larger sample of repositories would provide a fuller picture, as would comparable studies in other disciplines. Finally, some of the terms we used in the survey may not have been understood the same way by all respondents, as suggested by some of the free-text responses.

Our findings complement research into the factors that affect the choice of a researcher to archive their data at all. For example, Piwowar (2011) found that authors were most likely to archive the particular genomic datatype under study "if they had prior experience sharing or reusing data, if their study was published in an open access journal or a journal with a relatively strong data sharing policy, or if the study was funded by a large number of NIH grants. Authors of studies on cancer and human subjects were least likely to make their datasets available". It would be of interest in future studies to determine if disciplinary differences in the relative importance of the different factors affect both the willingness to archive data at all, and the choice of repository when the researcher decides to archive. Such joint analysis could be of great value in helping to customize disciplinary repositories to their communities of interest.

Acknowledgements

The authors thank William Piel, Hilmar Lapp, Michael Jones and Elena Feinstein, who provided assistance in collecting contact information for repository users. The data is available from the Dryad Digital Repository at doi:10.5061/dryad.51vs3. We acknowledge the financial support of the National Science Foundation (DBI-1147166).

Conflicts of interest

Shea Swauger was previously employed by the Dryad Digital Repository project and Todd J. Vision serves as Principal Investigator for the project and sits on its volunteer Board of Directors. Todd J. Vision is also the Associate Director for Informatics at the National Evolutionary Synthesis Center, where TreeBASE is currently hosted.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., & Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology 215*(3), 403–410. doi:10.1016/S0022-2836(05)80360-2

Berkley, C., Bowers, S., Jones, M.B., Madin, J.S., & Schildhauer, M. (2009). Improving data discovery for metadata repositories through semantic search. In *CISIS'09. International Conference on Complex, Intelligent and Software Intensive Systems.* (pp. 1152–1159). IEEE. doi:10.1109/CISIS.2009.122

Bone, D., & Burns, P. (2011). An overview of content archiving services in scholarly publishing. Retrieved from http://allenpress.com/system/files/pdfs/library/archiving-whitepaper.pdf

Brokhoff, P.B., Best, D.J., Rayner, J.C.W. (2003). Using Anderson's Statistic to compare distributions of consumer preference rankings. *Journal of Sensory Studies 18*, 77–82. doi:10.1111/j.1745-459X.2003.tb00374.x

Callaghan, S., Tedds, J., Kunze J., Khodiyar, V., Lawrence, R., Mayernik M., Murphy F., Roberts T., & Whyte, A. (2014). Guidelines on recommending data repositories as partners in publishing research data. *International Journal of Data Curation 9*(1), 152–163. doi:10.2218/ijdc.v9i1.309

Campbell, E.G., Clarridge, B.R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N. A., & Blumenthal, D. (2002). Data withholding in academic genetics. *JAMA: The Journal of the American Medical Association, 287*(4), 473–480. doi:10.1001/jama.287.4.473

Catling, J.C., Mason, V.L., & Upton, D. (2009). Quality is in the eye of the beholder? An evaluation of impact factors and perception of journal prestige in the UK. *Scientometrics*, *81*(2), 333–345.

Clayton, N., Biddle, R., & Tempero, E. (2000). A study of usability of web-based software repositories. Proceedings of the International Conference on Software Methods and Tools (SMT) 2000: 51–58. doi:10.1109/SWMT.2000.890420

CODATA-ICSTI Task Group on Data Citation Standards and Practices. (2013). Out of cite, out of Mind: The current state of practice, policy, and technology for the citation of data. *Data Science Journal 12*, CIDCR1–CIDCR75. doi:10.2481/dsj.OSOM13-043

Costello, M.J., Michener, W.K., Gahegan, M., Zhang, Z.Q., & Bourne, P.E. (2013). Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology and Evolution 28*(8) 454–461. doi:10.1016/j.tree.2013.05.002

Edmunds, S.C., Pollard, T.J., Hole, B., & Basford, A.T. (2012). Adventures in data citation: sorghum genome data exemplifies the new gold standard. *BMC Research Notes, 5*(1), 223. doi:10.1186/1756-0500-5-223

Fegraus, E.H., Andelman, S., Jones, M. B., & Schildhauer, M. (2005). Maximizing the value of ecological data with structured metadata: An introduction to ecological metadata language (EML) and principles for metadata creation. *Bulletin of the Ecological Society of America, 86*(3), 158–168. doi:10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2

Fried Foster, N., & Gibbons, S. (2005). Understanding faculty to improve content recruitment for institutional repositories. *D-Lib Magazine 11*(1). Retrieved from http://www.dlib.org/dlib/january05/foster/01foster.html

Goovaerts, M., & Leinders, D. (2012). Metadata quality evaluation of a repository based on a sample technique. In *Metadata and Semantics Research* (pp. 181–189). Springer Berlin Heidelberg.

International Organization for Standardization. (2012). ISO 16363:2012 Space data and information transfer systems – Audit and certification of trustworthy digital repositories. Geneva, Switzerland: ISO.

Jacobs, J.A., & Winslow, S.E. (2004). Overworked faculty: Job stresses and family demands. *The Annals of the American Academy of Political and Social Science, 596*(1), 104–129. doi:10.1177/0002716204268185

Jones, S. (2012). Developments in research funder data policy. *International Journal of Digital Curation 7*(1), 114–125. doi:10.2218/ijdc.v7i1.219

Magee, A.F., May, M.R., Moore, B.R. (2014). The dawn of open access to phylogenetic data. *PLOS ONE 9*(10), e110268. doi:10.1371/journal.pone.0110268

Piwowar, H.A. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PLOS ONE, 6*(7), e18657. doi:10.1371/journal.pone.0018657

Rice, R., Ekmekcioglu, C., Haywood, J., Jones, S., Lewis, S., Macdonald, S., & Weir, T. (2013). Implementing the research data management policy: University of Edinburgh roadmap. *International Journal of Data Curation 8*(2), 194–204. doi:10.2218/ijdc.v8i2.283

Roche, D.G., Lanfear, R., Binning, S.A., Haff, T.M., Schwanz, L.E., et al. (2014). Troubleshooting public data archiving: Suggestions *to increase participation. PLOS Biology 12*(1), e1001779. doi:10.1371/journal.pbio.1001779

Sayogo, D.S., & Pardo, T.A. (2013). Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. *Government Information Quarterly 30*, Supplement 1, S19–S31. doi:10.1016/j.giq.2012.06.011

Stoltzfus, A., O'Meara, B., Whitacre, J., Mounce, R., Gillespie, E., Kumar, S., … & Vos, R. (2012). Sharing and re-use of phylogenetic trees (and associated data) to facilitate synthesis. *BMC Research Notes, 5*(1), 574. doi:10.1186/1756-0500-5-574

Whitlock, M.C. (2011). Data archiving in ecology and evolution: Best practices. *Trends in Ecology and Evolution, 26*(2), 61–65. doi:10.1016/j.tree.2010.11.006

Wicherts, J.M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLOS ONE, 6*(11), e26828. doi:10.1371/journal.pone.0026828